Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit

J. Dairy Sci. 93:3320–3330
doi:10.3168/jds.2009-2845

Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit

P. R. Amer*[1] and G. Banos†
*AbacusBio Limited, PO Box 5585, Dunedin, New Zealand
†Department of Animal Production, Faculty of Veterinary Medicine, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece

ABSTRACT

The aim of this study was to evaluate and quantify the importance of avoiding overlap between training and testing subsets of data when evaluating the effectiveness of predictions of genetic merit based on genetic markers. Genomic selection holds great potential for increasing the accuracy of selection in young bulls and is likely to lead quickly to more widespread use of these young bulls with a shorter generation interval and faster genetic improvement. Practical implementations of genomic selection in dairy cattle commonly involve results of national genetic evaluations being used as the dependent variable to evaluate the predictive ability of genetic markers. Selection index theory was used to demonstrate how ignoring correlations among errors of prediction between animals in training and testing sets could result in overestimates of accuracy of genomic predictions. Correlations among errors of prediction occur when estimates of genetic merit of training animals used in prediction are taken from the same genetic evaluation as estimates for validation of animals. Selection index theory was used to show a substantial degree of error correlation when animals used for testing genomic predictions are progeny of training animals, when heritability is low, and when the number of recorded progeny for both training and testing animals is low. Even when training involves a dependent variable that is not influenced by the progeny records of testing animals (i.e., historic proofs), error correlations can still result from records of relatives of training animals contributing to both the historic proofs and the predictions of genetic merit of testing animals. A simple simulation was used to show how an error correlation could result in spurious confirmation of predictive ability that was overestimated in the training population because of ascertainment bias. Development of a method of testing genomic selection predictions that allows unbiased testing when training and testing variables are estimated breeding values from the same genetic evaluation would simplify training and testing of genomic predictions. In the meantime, a 4-step approach for separating records used for training from those used for testing after correction of fixed effects is suggested when use of progeny averages of adjusted records (e.g., daughter yield deviations) would result in inefficient use of the information available in the data.
Key words: genomic selection, validation, selection index theory, simulation

INTRODUCTION

Genomic predictions of animal genetic merit offer several exciting but challenging opportunities for genetic improvement of livestock. The idea of genomic selection is likely to be particularly beneficial to dairy improvement schemes (e.g., Schaeffer, 2006). Accurate predictions of genetic merit of selection candidates come very late in the lives of semen-producing bulls, which contribute the majority of genetic progress in dairy industries throughout the world. Genomic selection facilitates wider use of younger bulls and potentially a less formalized progeny-testing structure (König et al., 2009).

The enhancement of phenotypic records as predictors of the genetic merit of selection candidates using genetic markers was first addressed in the context of selection index theory by Neimann-Sorenson and Robertson (1961). Much research effort on genetic markers has resulted, often with disappointing outcomes in terms of practical improvements to livestock breeding programs (Dekkers, 2004). Initial hopes that variation in traits inherited in a quantitative manner would be caused largely by independently acting genes with additive modes of inheritance have not been borne out (Visscher et al., 2008). Applications of genetic markers targeting traits with highly polygenic inheritance have also been proposed. Nejati-Javaremi et al. (1997) suggested that DNA markers could be used to improve the accuracy of genetic evaluations by replacing the pedigree-based numerator relationship matrix used in genetic evaluations with a relationship matrix based on markers such

Received October 21, 2009.
Accepted March 14, 2010.
[1]Corresponding author: pamer@abacusbio.co.nz

that animals carrying more similar variants of markers on average are assumed more closely related than those that do not. Meuwissen et al. (2001) evaluated methods for genetic evaluation that use very large numbers of markers and termed their use in breeding programs as genomic selection. Their simulations suggested that methods that give greater emphasis to markers with larger estimated effects resulted in more accurate genomic predictions.

Among the challenges for implementing genomic selection is the statistical issue of estimating the effects of large numbers of genetic markers available relative to the number of animals with both genotypes and phenotypic records. A consequence is that it is common to split data sets into training and testing subsets to evaluate the accuracy of and bias in genomic predictions before industry application. A further challenge revolves around the computing complexity of incorporating genomic information into conventional genetic evaluation processes in a routine way that is also robust. Practical solutions are rapidly being developed to overcome these challenges (e.g., VanRaden, 2008). These practical solutions use outputs from industry genetic evaluations to train predictions of genetic merit using markers, and then blend or combine the genomic predictions with conventional genetic evaluations using regression theory. Initial results for dairy genetic evaluation systems appear promising and so they are having a major effect on dairy genetic evaluation outputs and breeding strategies in several countries (e.g., Spelman et al., 2007; VanRaden et al., 2009).

Training and testing data sets should not overlap when verifying the predictive ability (both accuracy and bias) of any method proposed for genomic selection. However, national genetic evaluation systems are extremely complex, and it is often tempting to use predictions of genetic merit for both training and testing that come from a single national genetic evaluation run.

In this paper, the effect of using overlapping data sets for training and testing was evaluated when assessing the efficacy of genomic selection for increasing rates of genetic progress. A description is first provided of how the (realized) accuracy of genomic predictions of genetic merit using genetic markers can be inferred from their correlations with conventionally estimated breeding values under certain assumptions. Selection index theory was then used to quantify how overprediction of the realized accuracy of genomic selection, which occurs when overlapping data contribute to sire breeding values used in both training and testing, is affected by numbers of progeny and trait heritability. Finally, a simple simulation was used to demonstrate how overlapping data sets used for training and testing breeding values can lead to spurious relationships because of ascertainment bias being carried through into apparent predictive ability during testing.

MATERIALS AND METHODS

If phenotypes can be accurately precorrected for systematic environmental effects, EBV can be derived using selection index theory as first proposed by Smith (1936). Best linear unbiased prediction (Henderson, 1973) is a statistical method that simultaneously adjusts for known systematic environmental effects while predicting genetic merit using the principles of selection index theory. Both approaches are capable of using information from relatives in their prediction of genetic merit (Legates and Lush, 1954; Henderson, 1975). Predictions of genetic merit using genetic markers are much less subject to the systematic environmental biases that affect phenotypic records and so are highly amenable to selection index applications.

In the first two components of this study, selection index theory is used. First, to provide a theoretical basis for computing realized accuracies from genomic predictions in a testing data set, and second, to quantify the lack of independence of errors that can occur between training and testing animals when relatives of both are included in genetic evaluations contributing to either the training or the testing process. In the third component of this study, simulation is used to demonstrate how falsely inflated realized accuracies can be generated when training and testing have predictions of genetic merit that are not based on independent data sets.

Realized Accuracy of Molecular Predictions

Let **M** and **E** denote a molecular breeding value prediction and a conventional EBV prediction of genetic merit for an individual in a testing population, respectively. Molecular breeding value predictions (M) are based on SNP effect solutions. They correspond to the direct genomic values and should not be confused with what many refer to as a genomic or genomically enhanced breeding value, which is a prediction that combines M and E into a single number. Records from many individuals in a training population will have been used to inform the prediction of M. Let T denote the true additive breeding value for the same individual in a testing population; this is the value that M and E are attempting to predict. The accuracies with which the conventional EBV E predicts the true breeding value T is $r_{E,T}$. This will be available from the genetic evaluation system used to compute E. The accuracy with which the estimated molecular breeding value M

predicts the true breeding value T is $r_{M,T}$. This is an unknown statistic to be derived. The correlation between conventional EBV E and molecular breeding values M is $r_{E,M}$ and can be computed using the conventional formula to get the correlation between 2 variates.

Let ε_E denote the nonadditive genetic and environmental deviation of the prediction of genetic merit using an EBV (E) from a phenotype (P), with a mean equal to zero and variance denoted $\sigma^2_{\varepsilon E}$, such that E = ${r_{E,T}}^2$ (T + ε_E). The term ε_E will be henceforth referred to as prediction deviation. The variance $\sigma^2_{\varepsilon E}$ of the deviation term ε_E is related to the conventional mixed-model BLUP definition of prediction error variance. This is a more general form of the example under mass selection that $\hat{A} = h^2 \cdot P$, where $\hat{A}$ is a prediction of an animal's genetic merit from its own record (P), for a trait of heritability h^2. For this simple situation, h^2 is also the squared accuracy of prediction corresponding to the proportion of variance in A explained by P, just as ${r_{E,T}}^2$ is the proportion of variance in (T + ε_E) explained by E.

Similarly, the prediction deviation, ε_M, for a molecular breeding value has a mean equal to zero and variance denoted $\sigma^2_{\varepsilon M}$, such that M = ${r_{M,T}}^2$ (T + ε_M).

The correlation of conventional EBV (E) with molecular breeding values (M) is derived from

$$r_{E,M} = \frac{\mathrm{cov}(E,M)}{\sqrt{\mathrm{var}(E)\cdot \mathrm{var}(M)}}.$$

From expectations, this can be rewritten as

$$r_{E,M} = \frac{r^2_{E,T} r^2_{M,T}\left(\sigma^2_T + \sigma_{\varepsilon E,\varepsilon M}\right)}{\sqrt{r^2_{E,T}\sigma^2_T \cdot r^2_{M,T}\sigma^2_T}},$$

where σ^2_T is the genetic variance of the trait and $\sigma_{\varepsilon E,\varepsilon M}$ is the covariance between the errors in the predictors of E and M, assuming no covariance between T and ε_E, and T and ε_M. This equation simplifies to

$$r_{E,M} = r_{M,T}\left(r_{E,T} + \frac{r_{E,T} r_{\varepsilon E,\varepsilon M}\ \sigma_{\varepsilon E}\ \sigma_{\varepsilon M}}{\sigma^2_T}\right),$$

where $r_{\varepsilon E,\varepsilon M}$ is the correlation between prediction deviations for E and M, and $\sigma_{\varepsilon E}$ and $\sigma_{\varepsilon M}$are the standard deviations of these deviations, respectively. The correlation $r_{\varepsilon E,\varepsilon M}$ can also be interpreted as the correlation among prediction errors from standard mixed-model predictions of E and M scaled upwards to account for prediction accuracies by dividing the prediction error correlation by the product of $r_{E,T}$ and $r_{M,T}$.

Because it is useful to know $r_{M,T}$, the above formula can be rearranged and further simplified to obtain

$$r_{M,T} = \frac{r_{E,M}}{r_{E,T}\left(1 + \dfrac{r_{\varepsilon E,\varepsilon M}\sigma_{\varepsilon E}\sigma_{\varepsilon M}}{\sigma^2_T}\right)},$$

and in the situation where the deviations from phenotype for the prediction of E and M are uncorrelated (implying that the prediction errors as defined in standard mixed-model BLUP methodology are also uncorrelated), this simplifies further to

$$r_{M,T} = \frac{r_{E,M}}{r_{E,T}}.$$

In simple terms, the estimated accuracy of molecular breeding values can be calculated by scaling up the correlation between conventional EBV and molecular breeding values to account for the inaccuracy with which the conventional EBV reflect true breeding values. However, it is clearly not appropriate to use this last formula when there are correlations between errors for EBV and molecular breeding values; that is, $r_{\varepsilon E,\varepsilon M} \neq 0$. Ignoring positive correlations between E and M will inflate estimates of $r_{M,T}$; for example, if the covariance between E and M $\left(r_{\varepsilon E,\varepsilon M}\sigma_{\varepsilon E}\sigma_{\varepsilon M}\right)$ was 25% of the true genetic variance σ^2_T, then estimates of $r_{M,T}$ using $r_{E,M}/r_{E,T}$ will be biased upwards by 25%.

Correlations Between Errors

Several approaches for developing genomic predictions of genetic merit consider use of conventional EBV to develop prediction equations from molecular genetic markers (e.g., Stricker et al., 2009; Verbyla et al., 2009; Villumsen et al., 2009). In the section above, the situation of using EBV to test the accuracy of genomic predictors of genetic merit was investigated. Using EBV in genomic prediction development adds considerable genotype cost-savings in situations when only sires are genotyped and these sires have several recorded progeny. In these situations, attempts are normally made to create some independence between the information from training populations (where equations for genomic predictions are developed) and testing populations

(where the predictive ability of the genomic predictions is evaluated). Division of the population into training and testing subsets is usually required. This can be done horizontally through the pedigree tree (all animals born before particular years are used for training and all animals born at or after the same year are used for testing) or vertically through the pedigree tree by dividing the population into genetic lines, breeds, or strains. Horizontal division of the pedigree is preferred in situations where the goal is to predict the genetic merit of new generations of selection candidates using a combination of genomic information and phenotype-based EBV of their ancestors. Vertical division is more appropriate when the goal is to evaluate the predictive ability of genomic breeding values across breeds and subpopulations that differ from the training set.

Horizontal division of the pedigree appears to be the approach of choice in dairy cattle breeding situations, but it also poses greatest risk of correlation between prediction deviations for M and E. This section illustrates and quantifies the likely magnitude of this correlation using selection index theory. For illustrative purposes, a training population of progeny-tested sires, and a testing population made up of 1 progeny-tested son per sire is considered. Each progeny tested sire has n_{sire} progeny with a single phenotype, and each son has n_{son} progeny with a single phenotype. Although this scenario is simplistic, it is likely that most test population sires will be an immediate descendant of a training population bull for most dairy cattle applications with horizontal division of sires into training and testing individuals.

If breeding values were to be estimated using progeny records independently for sires and sons, the selection index weights $\left(b_{ind}\right)$ to be applied to progeny averages are well known as

$$b_{ind} = \frac{0.5nh^2}{1+0.25h^2\left(n-1\right)},$$

where h^2 is the trait heritability, and n is the number of progeny of the sire.

If both the sire's and the son's progeny records are used jointly in the prediction of genetic merit of the sire, which is the norm in routine genetic evaluation systems, then the selection index weights to be applied to the sire and son progeny averages ($b_{dep.sire.sire}$ and $b_{dep.sire.son}$, respectively) are

$$b_{dep.sire} = \begin{bmatrix} b_{dep.sire.sire} \\ b_{dep.sire.son} \end{bmatrix} = \mathrm{P}^{-1}\mathrm{C}_{sire},$$

where P, the phenotypic variance-covariance matrix for progeny means, and C, the covariance matrix between progeny means and the sire's true breeding value, are as follows:

$$\mathrm{P} = \begin{bmatrix} \dfrac{1+0.25h^2\left(n_{sire}-1\right)}{n_{sire}} & 0.125h^2 \\ 0.125h^2 & \dfrac{1+0.25h^2\left(n_{son}-1\right)}{n_{son}} \end{bmatrix},$$

$$\text{and } \mathrm{C}_{sire} = \begin{bmatrix} 0.5h^2 \\ 0.25h^2 \end{bmatrix}.$$

Similarly, if both the sire's and the son's progeny are used jointly in the prediction of genetic merit of the son, then the selection index weights are

$$b_{dep.son} = \begin{bmatrix} b_{dep.son.sire} \\ b_{dep.son.son} \end{bmatrix} = \mathrm{P}^{-1}\mathrm{C}_{son},$$

$$\text{where } \mathrm{C}_{son} = \begin{bmatrix} 0.25h^2 \\ 0.5h^2 \end{bmatrix}.$$

The expectation of correlation between true breeding values for the sire and the son is 0.5 under a typical additive and infinitesimal model of inheritance. The expectation of the correlation between EBV of sires and their sons when they are estimated from progeny records independently is

$$r_{ind} = \frac{0.125h^2}{\sqrt{\dfrac{1+0.25h^2\left(n_{sire}-1\right)}{n_{sire}} \cdot \dfrac{1+0.25h^2\left(n_{son}-1\right)}{n_{son}}}}.$$

The expectation of the correlation between EBV of sires and their sons when they are estimated jointly using both sets of information as in BLUP is

$$r_{dep} = \frac{b_{dep.sire}{}' \mathrm{P} b_{dep.son}}{\sqrt{b_{dep.sire}{}' \mathrm{P} b_{dep.sire} b_{dep.son}{}' \mathrm{P} b_{dep.son}}}.$$

However, it is not valid to make comparisons between r_{ind} and r_{dep} as defined above to evaluate the error correlation between the 2 (sires and sons) sets of breeding values. This is because there is less information used to

predict both sire and son breeding values for r_{ind} than for r_{dep}. Therefore, a more valid comparison to identify the error correlation is to derive the correlation between breeding values when the same amount of information is available for the dependent versus the independent sets of breeding values. This can be done by deriving selection index equations assuming additional sets of sire and son progeny, with each set having n_{sire} progeny for sires and n_{son} progeny for sons. One set of the sire's progeny is used in the prediction of the sire's EBV, the other set is used in the prediction of the son's EBV. In the same way, 2 sets of the son's progeny are used to provide equivalent amounts of information but with independence between the sire and son. Using this approach changes the covariance between sire and son EBV to be

$$\text{Cov}(dep.sire, dep.son) = b_{dep.sire}{}' \text{Q} b_{dep.son},$$

where

$$\text{Q} = \begin{bmatrix} 0.25h^2 & 0.125h^2 \\ 0.125h^2 & 0.25h^2 \end{bmatrix}.$$

Because the amount of information contributing to the breeding values is the same as when progeny information is used jointly, a new independent correlation is derived as

$$r_{ind.eq} = \frac{b_{dep.sire}{}' \text{Q} b_{dep.son}}{\sqrt{b_{dep.sire}{}' \text{P} b_{dep.sire} b_{dep.son}{}' \text{P} b_{dep.son}}}.$$

The correlation between prediction deviations can then be computed as $\sqrt{r_{dep}{}^2 - r_{ind.eq}{}^2}$.

One option to eliminate the error correlation between training and testing animals is to truncate the data so that completely independent data sets are used in evaluation of training versus testing animals. One example of this is to restrict the data used in the training analysis to that available before any testing animal had meaningful progeny. This approach is advocated by VanRaden (2008), specifically to avoid part-whole correlations between training and testing data. Thus, for the simple sire-son example, the sire's EBV would become independent of the average performance of the son's progeny. However, it is often less practical to restrict the estimation of breeding values for the testing set to exclude data that contributed to the training process. If data are scarce, then historic information may be required to adjust correctly for fixed effects and selection. For example, removing historic performance records from a data set where contemporary groups do not have strong recent connectedness could create a confounding between contemporary group environmental variation and the EBV of family lines that occur most frequently in contemporary groups. Therefore, it would be useful to determine if making the training set EBV independent of testing set performance records, but not vice versa, might assist with reducing the error correlation shown above.

Using this approach changes the covariance between sire and son EBV to be

$$\text{Cov}(ind.sire, dep.son) = b_{dep.sire}{}' \text{R} b_{dep.son},$$

where

$$\text{R} = \begin{bmatrix} \dfrac{1 + 0.25h^2\left(n_{sire} - 1\right)}{n_{sire}} & 0.125h^2 \\ 0.125h^2 & 0.25h^2 \end{bmatrix}.$$

Because the amount of information contributing to the breeding values is the same as when progeny information is used jointly, a new correlation is estimated between sire and son E, when sire E are estimated first without progeny of sons, and then sons E are estimated later including both sire and son progeny records, as

$$r_{indsire.eq} = \frac{b_{dep.sire}{}' \text{R} b_{dep.son}}{\sqrt{b_{dep.sire}{}' \text{P} b_{dep.sire} b_{dep.son}{}' \text{P} b_{dep.son}}},$$

and the correlation between prediction deviations is computed as $\sqrt{r_{indsire.eq}{}^2 - r_{ind.eq}{}^2}$.

The above equations for correlations between sire and son EBV were used to generate plots to show how the correlations between sire and son E are influenced by numbers of progeny per sire from 0 to 500 and trait heritabilities of 0.05, 0.2, and 0.4. The situation where sire E are predicted including son progeny records and vice versa were first plotted. Second, the situation was considered where sons E are predicted including sire records, but not vice versa. A situation where sons E are predicted including sire records but not vice versa, where sons have half as many progeny as sires is also plotted.

Simulation Showing Ascertainment Bias Confirmation

Examples of how errors in a training population of sires can be validated falsely when the testing population consists of sons with breeding values jointly estimated with training sire breeding values were demonstrated using simple simulations. There were 1,000 sire-son pairs simulated with 20 or 100 progeny recorded for each, for a quantitative polygenic trait with heritability of 0.05, 0.2, or 0.4. Sires were assumed unrelated. For each sire and a single dam per sire, 2,000 independent SNP genotypes were simulated with gene frequencies chosen randomly with equal chance to take a value between 0.05 and 0.5. Son genotypes were determined by randomly selecting an allele from each of the sire and dam. Sire and son breeding values were estimated either jointly from all progeny records available using selection index theory or from their own progeny records only. There was no linkage relationship between recorded trait genotypes and SNP.

Approximate molecular predictions were developed by regressing either sire EBV or sire progeny averages on each SNP individually using standard least squares regression. Although it is more common for much more complex multiple regression methods to be used in practice, independent regressions used here were suitable given the underlying simulation model and the objective of the simulation, which was to generate ascertainment bias in the molecular predictions. Resulting coefficients (b) were regressed according to their standard errors (SE) to obtain more conservative estimates of SNP effects (b^*) using the regression method of B. P. Kinghorn (University of New England, Armidale, Australia; personal communication):

$$b^* = b\frac{b^2 - SE^2}{b^2}.$$

The best 100 out of 2,000 SNP effects (b^*) were then selected to obtain those with the greatest absolute values, and these subsets were used to compute molecular breeding values for sons. Apparent realized accuracies of the molecular breeding values were taken as the correlation between son molecular breeding values and son EBV, divided by the accuracy for the son EBV. Correlations between molecular breeding values and the true breeding values of the sons were computed directly from the results. Each simulation was repeated 50 times, and a standard error of the replicated results was computed as the standard deviation of results divided by $\sqrt{50}$.

RESULTS

Figure 1 plots the expectations from selection index theory of the correlations between sire and son conventional EBV as the number of progeny per sire and per son increases equally. At low to moderate numbers of progeny per sire and son, and particularly for low heritability traits, the correlation between EBV estimated jointly is much higher than the expected correlation when there is no correlation of errors.

Figure 2 plots the expectations of the correlations between sire and son E when the sires E are estimated from a data set that does not contain the son's progeny. The difference between the higher and lower lines is reduced compared with that in Figure 1, because the error correlations come only through the contribution of the sire's progeny to the son's E.

Figure 3 plots the correlations when the sires E are estimated from a data set that does not contain the son's progeny but when sons have fewer progeny than the sires by a factor of 0.5. The biases increase again and at less than 40 progeny per sire are approximately midway between the biases shown for the 2 previous situations investigated (joint evaluation and sires evaluated independently with sires and sons having equal numbers of progeny).

Figure 4 plots the implied correlations between prediction deviations for sire and son E for the 3 situations shown in Figures 1 to 3. Only the 0.05 (highest line in each pair) and 0.2 (lowest line in each pair) heritability traits are plotted to avoid further clutter to the lower left-hand corner of the figure. For low to moderate heritability traits, the correlations are high at low to medium numbers of progeny, and moderate even with quite large numbers of progeny. When progeny of sons (testing) are not used in the predictions of genetic merit used to train sires, the magnitude of the correlations declines, but to a lesser degree if the sons (testing) have significantly fewer progeny than the sires (training), which is likely to occur in practice.

Results from the simulation in which high levels of ascertainment bias were expected are in Table 1. When training and testing data were independent, apparent realized accuracies of molecular predictions were low but increased slightly with higher heritability and higher numbers of progeny of training sires and testing sons. These low values reflect the lack of linkage associations between SNP markers and the completely polygenic simulated trait. The markers do have some true predictive ability, with correlations between son genomic breeding values and the son true breeding values significantly greater than 1. This is because higher genetic merit sires will tend to have both higher marker scores and higher genetic merit sons. At the same time,

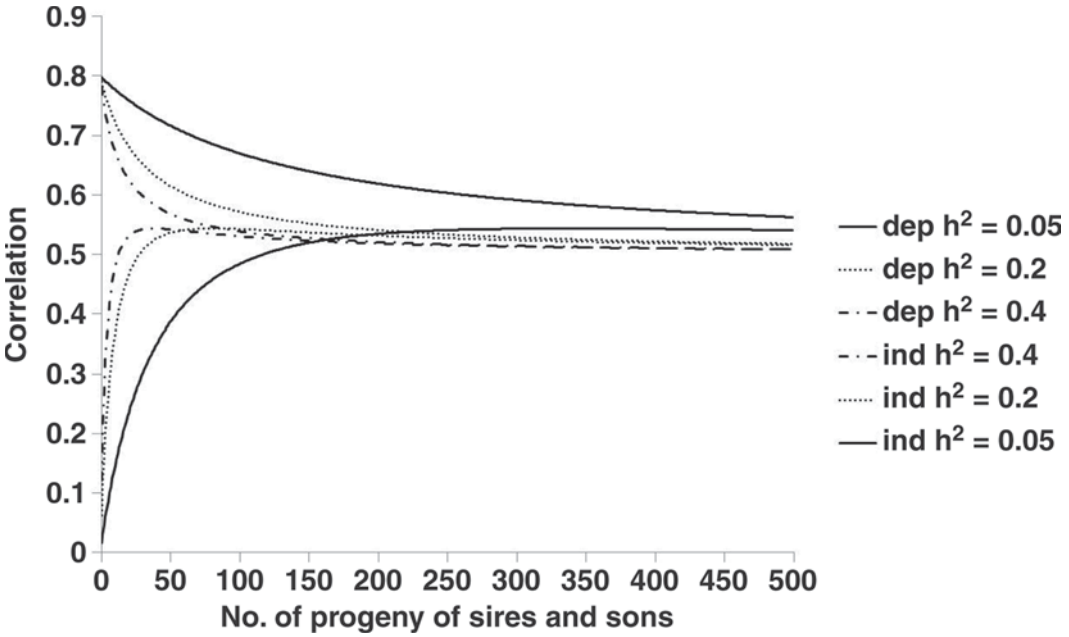

Figure 1. Expectations of correlations between progeny-tested sire and progeny-tested son breeding values estimated dependently (dep; higher lines) and independently (ind; lower lines) as the number of progeny per sire and son increase and with 3 different trait heritabilities (h^2).

sires and sons tend to inherit the same markers. Other studies have identified predictive ability of markers either under an approximately infinitesimal model or in the absence of linkage disequilibrium between markers and quantitative trait loci (Villanueva et al., 2005; Habier et al., 2007; Hayes and Goddard, 2008). For low heritability traits, and with low numbers of progeny, the apparent realized accuracies were grossly inflated when sires' EBV were not independent of sons' progeny records and vice versa. Even when sires' EBV did not

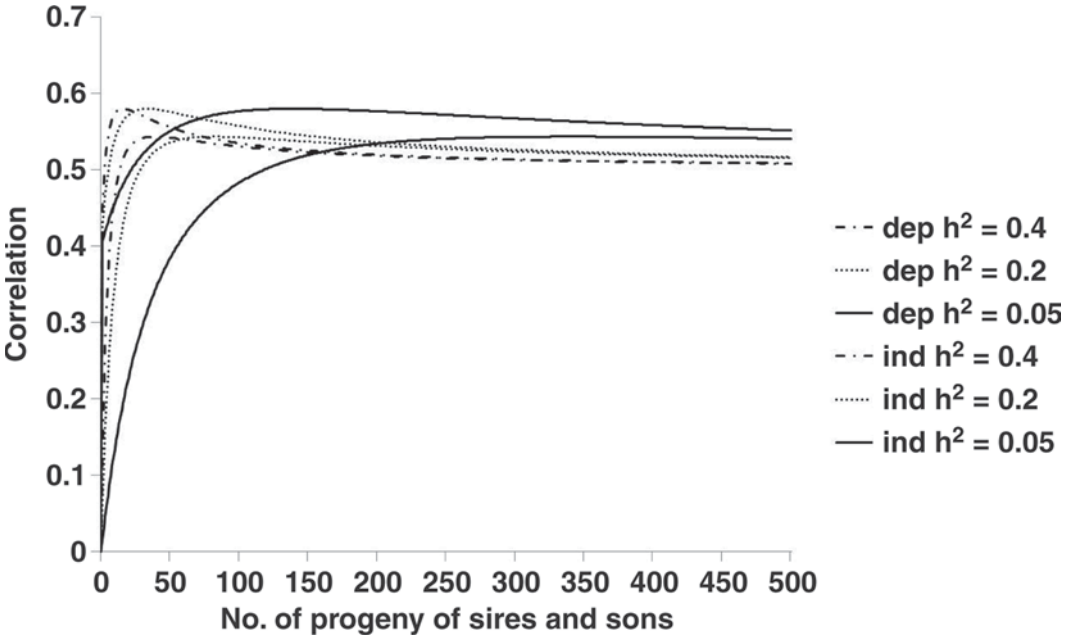

Figure 2. Expectations of correlations between progeny-tested sire and progeny-tested son breeding values when sire breeding values are estimated independently but son breeding values use information from their sire's progeny (dep; higher lines) and when both sire and son breeding values are estimated independently (ind; lower lines) as the number of progeny per sire and son increase and with 3 different trait heritabilities (h^2).

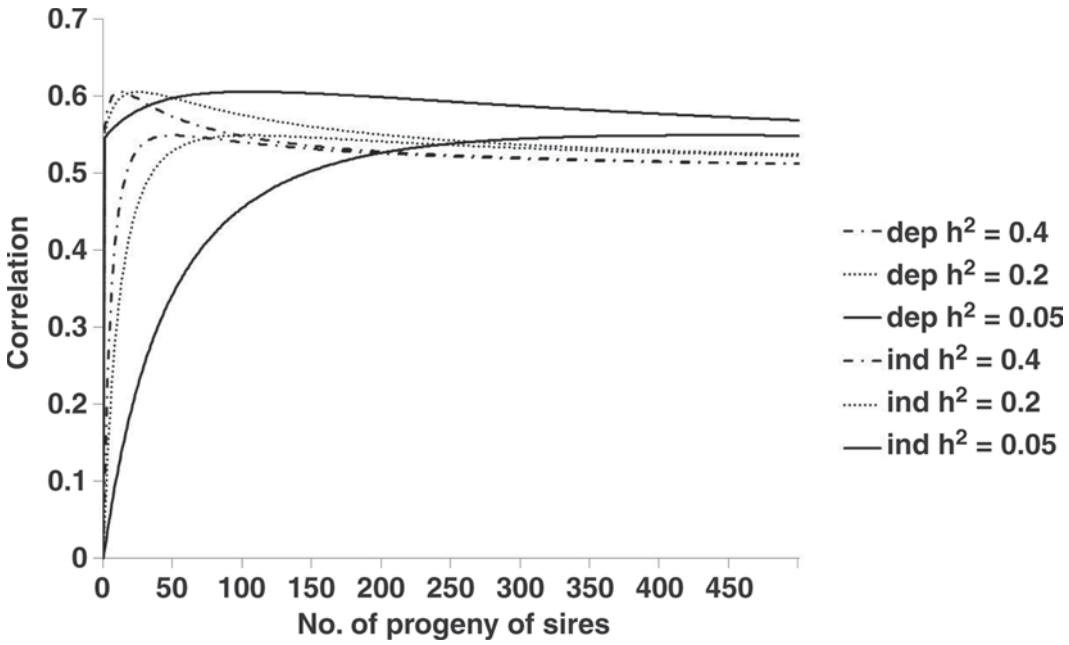

Figure 3. Expectations of correlations between progeny-tested sire and progeny-tested son breeding values when sons have one-half as many progeny as sires, sire breeding values are estimated independently but son breeding values use information from their sire's progeny (dep; higher lines), and when both sire and son breeding values are estimated independently (ind; lower lines) as the number of progeny per sire increase and with 3 different trait heritabilities (h^2).

incorporate sons' progeny records, the apparent realized accuracies were still significantly inflated for low to moderate heritabilities and low to moderate numbers of progeny records. Some of the increase in accuracy can be attributed to more accurate sire EBV when the progeny information of their sons is included. As shown by Habier et al. (2007), even in the absence of linkage disequilibrium, the markers capture relationship information, and so the predictive ability of the markers in sons increases as the accuracy of the sire EBV increases. However, these effects are a very small proportion of the many-fold increases in apparent accuracy shown in Table 1.

DISCUSSION

This paper highlights a potential problem with using EBV from national genetic evaluations to both train and test molecular predictions of genetic merit using DNA markers. Correlations between prediction deviations for molecular and conventional predictors of genetic merit cause inflated estimates of realized accuracies of molecular breeding values. The simulations here show that a genomic predictor could be developed that has a reasonable correlation with the jointly estimated breeding values of sires, and even if this predictor had minimal linkage association with true genetic merit, it could be mistaken to have moderate to high predictive accuracy in the sons. This is because the correlation between sire and son EBV from a single genetic evaluation can be much higher than the correlation between their true breeding values. The genomic prediction would have a correlation close to 0.5 between sires and sons, assuming the heritability for the genomic predictor was close to 1.

Although the focus of this study has been on realized accuracies for molecular predictions, the same conclusions will apply when evaluating realized accuracies when molecular information is integrated with conventional breeding values to obtain a final and integrated prediction of merit (sometimes termed a molecular breeding value). When molecular estimates of breeding values and their accuracies are overestimated, too much emphasis will be given to marker information when it is being combined with conventional estimates of genetic merit. Our results therefore have implications for both molecular (genetic marker information only) and genomic (genetic marker information combined with a selection candidates conventional estimate of genetic merit) predictions.

The problem is greatest when heritability is low and there are low numbers of progeny per sire. Under these circumstances, the error correlation is greatest for training animals that are closely related to testing animals. In principle, a correction could be made to realized accuracies to account for this, but the simplistic situa-

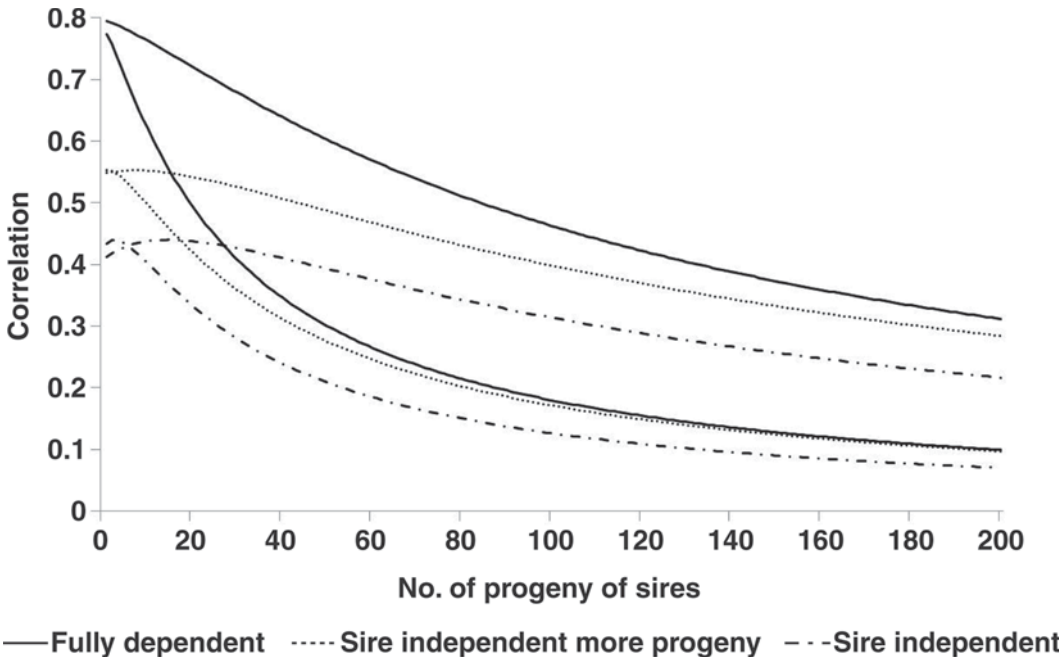

Figure 4. Differences between expectations of correlations between progeny-tested sire and progeny-tested son EBV where the same data contribute to both sire (training) and son (testing) EBV (fully dependent), son EBV use sires' progeny data but not vice versa, and sires have twice as many progeny as sons (sire independent more progeny), and son EBV use sires' progeny data but not vice versa and sires have the same number of progeny as sons (sire independent). Higher and lower lines in each pair correspond to heritabilities of 0.05 and 0.2, respectively.

tion modeled here of training animals that are sires of testing animals is practically unlikely, and the actual relationships and prediction errors among training and testing animals are likely to be much more complex. A more sensible approach is therefore to ensure that predictions of genetic merit for animals used in training do not use records of close relatives of animals used for testing and vice versa.

VanRaden (2008) and VanRaden et al. (2009) describe and evaluate methods that circumvent the problem of training and testing dependence through the use of daughter yield deviations as defined by VanRaden and Wiggans (1991). Nevertheless, deregressed evaluations are alluded to as potential response variables of interest in VanRaden (2008) and this could create some confusion, particularly as deregressed evaluations may be more conveniently obtained from genetic evaluation results than daughter yield deviations. Restricting response variates to corrected progeny averages may also lead to loss of other records that could contribute information without contributing substantially to prediction error correlations between training and validation animals. The performance records of genotyped animals with small numbers of progeny are a good example and would be lost if, for example, genotyped cows were used for training or validation. Some recent studies (Stricker et al., 2009; Verbyla et al., 2009) appear to prefer the use of deregressed evaluations over daughter yield deviations when evaluating genomic selection in dairy cattle. Villumsen et al. (2009) compared the use of daughter yield deviations and EBV as response variables, but did not acknowledge the potential problem of prediction error correlation as outlined here.

Development of a method that could correct for prediction error correlations between closely related animals across the training-testing boundary would be very useful for practical implementation of genomic selection. Attempts to apply genomic selection in situations where genotyped animals have small numbers of progeny recorded for traits with low heritability are likely to become increasingly prevalent. This is particularly the case for species with short generation intervals but in which phenotypic recording of high value low heritability traits is expensive. In the meantime, a 4-step genetic evaluation process that feeds the genomic selection training and testing process could be considered as follows:

- Step 1: conduct a full genetic evaluation capturing fixed effects solutions and adjust phenotypic records for fixed effects;
- Step 2: separate adjusted phenotypic records into those that can be used to contribute to training response variables and those that can be used to contribute to testing response variables;

Table 1. Apparent realized accuracies of genomic breeding values and correlations between genomic breeding values and the true breeding values of the sons (son true correlation) from simulations and their standard errors for traits with 3 heritabilities (0.05, 0.2, and 0.4) when both sires and sons have either 20 or 100 progeny with records, and training is based on sire breeding values with and without inclusion of records from sons' half-sib progeny and testing based on son breeding values with and without inclusion of records from sires' half-sib progeny

h^2	Progeny	Training	Testing	Apparent realized accuracy	SE	Son true correlation	SE
0.05	20	Sire only	Son only	0.048	0.006	0.056	0.004
0.05	20	Sire only	Sire/son	0.198	0.006	0.056	0.004
0.05	20	Sire/son	Sire/son	0.342	0.006	0.095	0.004
0.05	100	Sire only	Son only	0.087	0.005	0.091	0.003
0.05	100	Sire only	Sire/son	0.152	0.005	0.091	0.003
0.05	100	Sire/son	Sire/son	0.224	0.004	0.134	0.004
0.2	20	Sire only	Son only	0.084	0.006	0.093	0.005
0.2	20	Sire only	Sire/son	0.157	0.006	0.093	0.005
0.2	20	Sire/son	Sire/son	0.225	0.005	0.120	0.004
0.2	100	Sire only	Son only	0.105	0.004	0.104	0.004
0.2	100	Sire only	Sire/son	0.125	0.004	0.104	0.004
0.2	100	Sire/son	Sire/son	0.147	0.005	0.118	0.004
0.4	20	Sire only	Son only	0.105	0.005	0.101	0.005
0.4	20	Sire only	Sire/son	0.147	0.005	0.101	0.005
0.4	20	Sire/son	Sire/son	0.179	0.005	0.118	0.005
0.4	100	Sire only	Son only	0.124	0.005	0.123	0.005
0.4	100	Sire only	Sire/son	0.134	0.005	0.123	0.005
0.4	100	Sire/son	Sire/son	0.125	0.004	0.110	0.004

- Step 3: do a full animal model BLUP analysis with pedigree and phenotypes included for animals with phenotypic records designated for training, but with no fitting of fixed effects required, and use the resulting EBV after deregression as response variables for training; and
- Step 4: do a full animal model BLUP analysis with pedigree included for all animals (those with both training and testing records) but only using phenotypic records from animals designated for testing, ignoring fixed effects and use the resulting EBV after de-regression as response variables for testing. This method is likely to be less effective for traits whose genetic evaluations are compromised when historic selection effects are not accounted for.

CONCLUSIONS

Testing procedures to evaluate realized accuracies for genomic predictions of genetic merit can result in overconfidence for low heritability traits and when genotyped animals have low numbers of progeny, if care is not taken to ensure independence of prediction errors for related animals in the training and testing subsets of the data. Use of daughter yield deviations as the response variables for training and testing should overcome this problem, but it is not always simple and practical to use them, and furthermore, this approach may result in a significant loss of useful data, particularly when genotyped cows are to be used for either training or validation.

ACKNOWLEDGMENT

Peter Amer gratefully acknowledges financial support for this study from Ovita Limited (Dunedin, New Zealand).

REFERENCES

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313–E328.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Hayes, B. J., and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci. 86:2089–2092.

Henderson, C. R. 1973. Sire evaluation and genetic trend. Pages 10–41 in Animal Breeding Genetics Symposium in honor of Dr. J. L. Lush. American Society of Animal Science and American Dairy Science Association, Champaign, IL.

Henderson, C. R. 1975. Use of all relatives in intraherd prediction breeding values and real producing abilities. J. Dairy Sci. 58:1910–1916.

König, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92:382–391.

Legates, J. E., and J. L. Lush. 1954. A selection index for fat production in dairy cattle utilizing the fat yields of the cow and her close relatives. J. Dairy Sci. 37:744–753.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745.

Niemann-Sorenson, A., and A. Robertson. 1961. The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agric. Scand. 11:163–196.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.

Smith, H. F. 1936. A discriminant function for plant selection. Ann. Eugen. 7:240–250.

Spelman, R. J., J. Arias, M. Keehan, V. Obolonkin, A. Winkelman, D. Johnson, and B. Harris. 2007. Application of genomic information in a dairy cattle breeding scheme. Proc. Assoc. Advmt. Anim. Breed. Genet. 17:471–478.

Stricker, C., J. Moll, H. Joerg, D. J. Garrick, and R. L. Fernando. 2009. First results on genome-wide genetic evaluation in Swiss dairy cattle. Page 303 in Book of abstracts, EAAP 60th Annual Meeting, Barcelona, Spain. EAAP, Rome, Italy.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Verbyla, K., P. Bowman, B. Hayes, H. Raadsma, M. Khatar, and M. E. Goddard. 2009. Comparison of Bayesian methods for genomic selection using real dairy data. Page 294 in Book of abstracts, EAAP 60th Annual Meeting, Barcelona, Spain. EAAP, Rome, Italy.

Villanueva, B., R. Pong-Wong, J. Fernández, and M. A. Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83:1747–1752.

Villumsen, T. M., L. Janss, P. Madsen, and M. S. Lund. 2009. EBV and DYD as response variable in genomic predictions. Page 299 in Book of abstracts, EAAP 60th Annual Meeting, Barcelona, Spain. EAAP, Rome, Italy.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era—Concepts and misconceptions. Nat. Rev. Genet. 9:255–266.